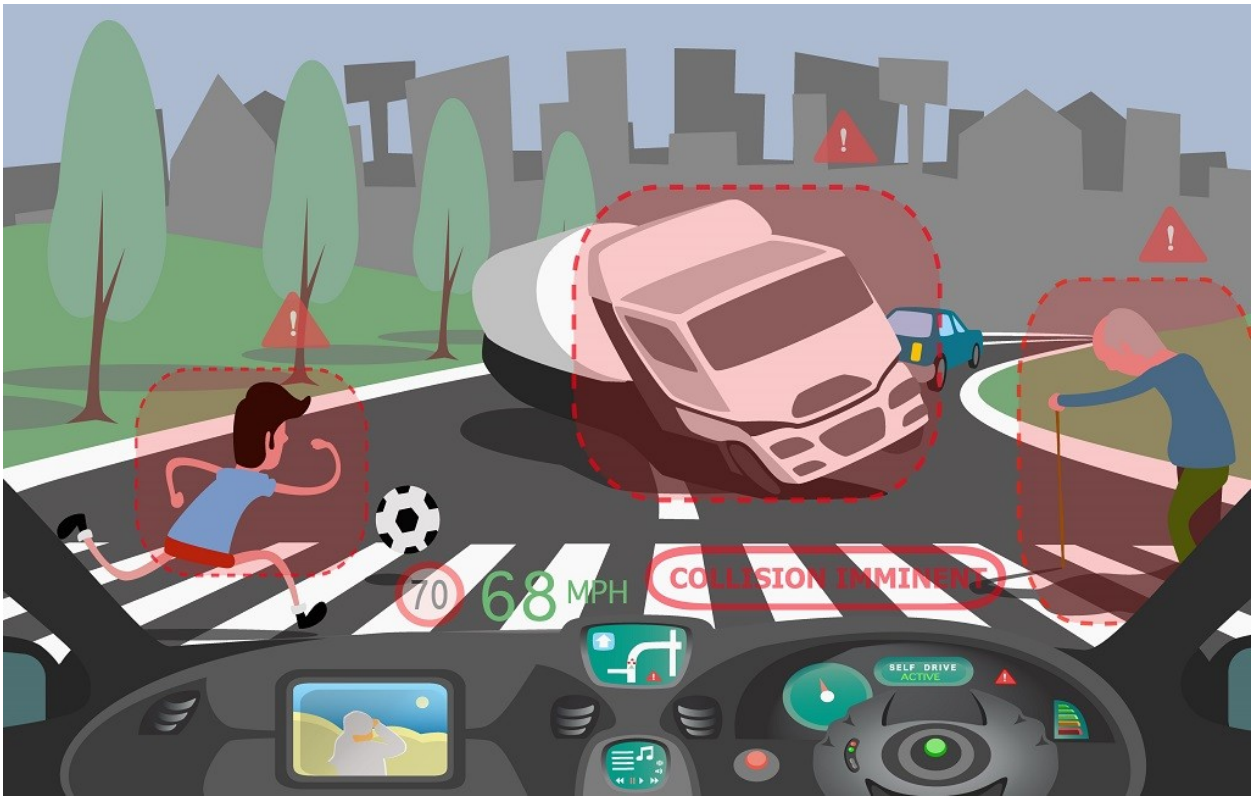
Self-driving cars, moral theory, and outsourcing responsibility

Part 3 of robot ethics [parts 1 (consciousness and suffering) and 2 (robot love)]



Never in the history of humanity have we allowed a machine to autonomously decide who should live and who should die, in a fraction of a second, without real-time supervision. We are going to cross that bridge any time now, and it will not happen in a distant theatre of military operations; it will happen in that most mundane aspect of our lives, everyday transportation. Before we allow our cars to make ethical decisions, we need to have a global conversation to express our preferences to the companies that will design moral algorithms, and to the policymakers that will regulate them.

~ "The Moral Machine experiment"[1]

"Being a homeless person is like jaywalking all the time."

~ Jean-Francois Bonnefon

With greater precision comes greater responsibility... There is no special reason to take what people say as evidence of anything but what people say.

1. "The Moral Machine experiment," Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon & Iyad Rahwan. *Nature*, volume 563 (2018), p. 63.

It's not like you are sitting in the back seat of an Uber. If the driver has an accident, it's not your fault, it is the driver's. It's like *this* car is yours. You chose it.[2] But it was sold to you with a software package that will take over the moment to moment decisions of driving armed with a vast store of procedures to follow in nearly any situation. It was designed to carry them out with a speed, precision, and freedom from distraction, unmatchable by any human like you. There is no doubt that *all of us* are safer because (highly fallible) *you* are *not* driving this car. The decisions the car makes will be preset to accord with a predetermined and vetted "uncontroversial" moral principle—one that *sounded good* to the car's makers.

Or maybe the car company will give you a choice of algorithm. You opt for software that instantiates values that seem most intuitive to you from among a publicly acceptable set. Suppose you choose a utilitarian car when you could have chosen a deontological car or a virtue or care ethics car. If so, you are still, then, in some sense, responsible for what your car does. But how is this going to work when different people choose different value systems to build into the AI of their cars?

Either way, whether we choose and therefore "own" the algorithm, or we outsource this responsibility to institutions or "experts," there are questions that need to be answered.

For the sake of this discussion, we are going to assume our self-driving cars will be able to follow their programmed values with unerring precision: they always perform as designed.[3] The issue is about their governing moral theory. From among the viable ones, which will it be? Let's further assume, that car makers and their legislative regulators are not going to invent moral theories out of whole cloth. That would be like re-inventing the wheel. They are going to look to people who know all about the "wheels" of moral theory to help decide which top level agenda the governing principle built into your car will serve. The problem is that there is no single top level moral theory surrounded by consensus. Rather, there are a handful. And, while 99% of the time they may agree, when push comes to crash they often don't. That little percentage is very real. Insurance companies know this well and can give it a dollar value.

Moral theories will have to get more fine-grained in a way they have never in the past needed to be. This is because fallibility is built into us. The extra precision of the theory would be wasted on us. *But not so with AI.* A *human* driver faced with the trilemma in the cartoon would be said

2. Or you let policy-makers choose it for you as in the case of AI controlled public transport.
3. The likelihood they might not is a technological concern and not particularly interesting from a moral theoretic point of view.

to be a “tragic” situation. No matter what action they take, it is hard to assign fault. Again, *not so with AI.* Some intelligence further back in the train of decision events, not subject to the felt urgency of the situation, would have to be saddled with responsibility. The programmers? The policy-makers? The social scientists who poll what ordinary people think? The moral theoreticians consulted?

The theory might be a variety of consequentialism, maybe *utilitarianism*. It might dictate that in all situations when the car cannot avoid harm altogether, it will take the action such that it causes the least amount of harm to all the entities vulnerable to harm involved.

The theory might be *deontological*. In which case, first, it is imperative that we show the utmost and equal respect for the well-being of all rational agents involved. The minimization of physical harm is not always the best way to do this. Since the class of those included among rational agents is far smaller than the class of sentient agents, and only the former demand direct moral consideration, it may be the case that some vulnerable to harm may be *worth* more than others. Complicating things further, rational agents often take on special duties to others thus elevating the moral status of these entitled others over that of other rational but not entitled agents. (Ex. a mother may have special duties to her child that she does not have to another’s child, or a spouse to their spouse and not to a third party.)

The theory may be some form of contractarianism, or virtue theory, or care ethics... And there are variations and hybrids of these plausible theories that may inform the algorithms of SD cars… *too many*, in fact. That’s a problem in itself: that there are too many. Add that they each have problems... You might think we should choose the one with the least problems. Which one is that? It depends on whom you ask.

What’s a car to do?

Simple: just ask lots of people, right?

A recent experiment tried to determine what most people think about what the right thing to do is in trolley problem moral dilemmas like the one in the cartoon. Before we get to the intriguing results, we will review what a trolley problem is and its relevance to ethical decision making.

Since Plato, philosophers have been doing thought experiments to ferret out moral intuitions from wherever they hide when we are not being stressed to think about them. They do this by subjecting them to “crash” tests with varying results. In recent decades, the trolley problem has

emerged as a tool for evaluating moral intuitions in dire situations. Originally, the brainchild of Philippa Foot, later developed by Judith Jarvis Thompson,[4] the trolley problem emerged from the early abortion debate. Now it has developed into a small industry of thought experimentation —"trolleyology," directed at testing top level ethical principles against our "gut level" moral intuitions. Trolley problems have found a new home in the context of autonomous vehicles—the form of AI, it seems likely, we will encounter sooner rather than later.

Suppose you are the driver of a trolley whose brakes have failed. You are headed down a track that branches in two. If you do nothing, the trolley will head down one track and kill five people. If you steer the trolley down the other track, you will kill one person. Foot deems the situation one where it is permissible to take the action that results in the death of one and saves the five. Whatever regard you owe the one, surely you owe five times that to the group.

But now suppose you are a surgeon who may save five people fatally in need of organ transplants by killing a sixth person and harvesting their organs to save the five. Despite the superficial similarity—killing one to save five, Foot thinks our moral intuitions are different between these two scenarios. The rule "whatever regard you owe the one, you owe five times that to the group of five" does not apply in this case. Why not? Because, Foot suggests, we have this moral qualm: there is difference between *allowing someone to die but not intending it* in order to save others, on the one hand, and *intending that someone die* in order to save others, on the other.

Another way to put this, is that letting die is different from intending death. Philippa Foot writes,

> The doctrine of double effect[5] offers us a way out of the difficulty, insisting that it is one thing to steer towards someone foreseeing that you will kill him and another to aim at his death as part of your plan.[6]

What is at the heart of the distinction suggested by the "doctrine of double effect"? It seems that we have stronger negative duties not to interfere in someone's life than we have positive duties to benefit them. Or, alternatively, that we have stronger negative rights not to be interfered with than we have positive rights to demand that others benefit us.

In the trolley case, it is the *negative* rights of five versus the *negative* rights of one. Five wins. We are comparing apples to apples.

4. Philippa Foot, "The Problem of Abortion and the Doctrine of the Double Effect," originally appeared in the *Oxford Review*, Number 5, 1967. Judith Jarvis Thomson, "Killing, Letting Die, and the Trolley Problem," 59 *The Monist*, 204-17, 1976.
5. From the SEP entry: "According to the principle of double effect, sometimes it is permissible to cause a harm as a side effect (or 'double effect') of bringing about a good result even though it would not be permissible to cause such a harm as a means to bringing about the same good end."
6. Foot, *op. cit.*

In the transplant case, the *negative* right of the potential (supposedly unwilling) organ donor trumps any amount of *positive* rights the five organ-needy have. These are qualitatively different fruits.

We owe others non-interference *more* than we owe them benefits. Again, first avoid harm, then, if possible, do good (to paraphrase the Hippocratic oath). The core idea here is the *primacy* of avoiding harm.

Is this primacy correct? Many agree with Foot that it is. But you may not…

Consider the use David Benatar's anti-natalism argument makes of this primacy idea: Benatar makes a powerful case that if we really want to do the right thing we need to become extinct as a species. His argument relies on a version of the doctrine that we owe people non-interference *before* we owe them benefit to conclude that it is immoral to bring new life into the world. We owe the world (all the people in it) *less* suffering if it is within our power to avoid it, while we are not under any obligation to risk more suffering by creating the possibility of enjoyment. It is no excuse to say that non-existents, the unborn, left unborn, would miss out on the good things life may offer them. Even leaving aside the strange thought that a *non-existent* could possibly miss out on anything! (Don't you have to be a player in a game to be in a position to lose?) Do we think we improve their lot by exposing non-existents to the *certain* suffering that attends anticipation of death? Even if they are lucky enough, against all odds, to have *that* anticipation and its event be the *only* suffering life brings them?

You may find Benatar's argument repellant. At least it should give you pause if you really do believe in things like the Hippocratic Oath[7] or "doctrine of double effect." Maybe doctors *should* risk doing harm in order to do good? Maybe physician-assisted suicide *should* be morally equivalent to assisted murder?

The point of thought experiments like the trolley problem is not to settle what to do in specific, highly improbable, dilemmic scenarios. It is to reveal our thinking in situations that test our gut feelings about morality and the rationalizations we concoct to defend those feelings.

So what are our gut feelings about morality? Many moral philosophers think that "feelings," at least as a basis for decisions about right and wrong, are highly overrated for all kinds of reasons. We'll return to discuss these reasons, but apparently some think that first affective reactions *are*

7. It's not clear that Hippocrates (460–370 BC) himself actually put it this way. "Practice two things in your dealings with disease: either help or do not harm the patient," is perhaps closer to the actual Hippocratic statement, which doesn't strictly rule out an inclusive disjunctive reading: that one may do both or that one may do one in doing the other. The primacy of avoiding harm, *primum non nocere*, is a later development. The principle of non-maleficence anchors Benatar's anti-natalist case.

relevant and, like any phenomena, their normative relevance can be the object of discovery by social, anthropological, and psychological sciences.

"The Moral Machine"

"The Moral Machine experiment" is the title of a recent groundbreaking study by a group of social scientists, part of an MIT project, recently published in *Nature*. It asks, is there consensus among people (untainted by moral theory) worldwide about the moral decisions that autonomous vehicles should be designed to make? The experimenters canvassed the opinions of millions of subjects from many parts of the world on their reactions to various trolley problems. Among the dimensions measured were whether factors like the age, sex, number, socioeconomic status, passenger/non-passenger status, law-abidingness, species, etc.[8] of the various parties to an accident involving an autonomous vehicle make a difference to how such vehicles *should* be programmed in the view of a "random" self-selected but very large and culturally diverse sample of "ordinary" people. The scale of the study was both facilitated and limited by the fact that it was conducted online. The idea was to take a "moral reading," so to speak, of how people felt about, for example, whether *if someone must die* should it be an 8 year old or an 80 year old, a female or a male, a doctor or a homeless person, one or five people, an animal or a human, a jaywalker or someone dutifully waiting for the "walk" sign to light up, a passenger in the self-driving car or a bystander on the sidewalk, and so forth.

While very interesting as a kind of moral anthropology, the study seems to presuppose a moral framework (specifically, a consequentialist one) that mars its attempt to be culturally objective. The study is likely going to get heavy criticism from thinkers inclined toward deontology or other theoretical orientations.

Moreover, even if most people agreed on a moral theory, *would that make it the correct one?* Moral realists, in particular (those who think there are rights and wrongs irrespective of whether anybody even knows about them, let alone agrees about them), are going to have a field day criticizing the methodological assumptions of this survey.

Suppose most people think the earth is flat, would that mean the earth really is flat? It is likely the case that a flat earth theory was held almost universally at sometime in the past, in the Greek world, during the time of the pre-Socratics, for example. (Note that Aristotle felt compelled to offer evidence that it was *not* flat.)

Suppose most people (including women) thought that women were intellectually and morally deficient, and so an education would be wasted on them, and so they should not own or inherit property or vote or hold public office. (This was the prevailing sentiment in the US and the UK,

8. Curiously, the study did not specifically ask about whether the racial or ethnic or foreign-looking aspects of a party to an accident should make a difference. It is naive to think that if some of the other parameters measured would factor in these decisions, that race, ethnicity, or nationality would not.

for example, well into the 19th Century except among a scarce set of privileged women. Why were these women scarce? Because most women were uneducated and intellectually deficient. Why were they that? Because… and so on.) Would "most people" then be right? Since when has what "most people" think been an indication of what is *objectively* true?[9]

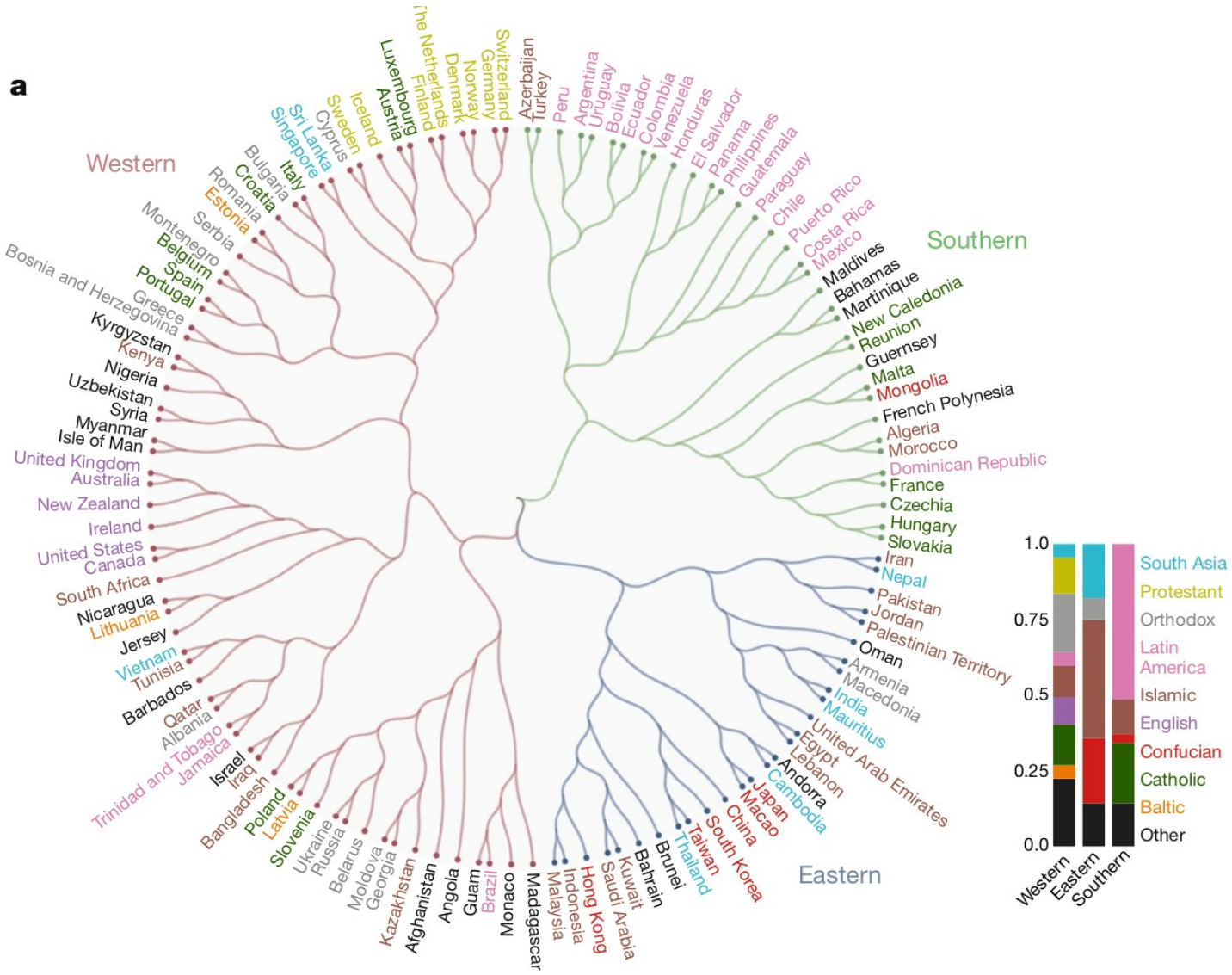
Perhaps it is true that, if *most* people were not *already* disturbed by an act of murder, we could never convince them that they were right in believing the act wrong. A moral theory cannot flatly oppose widely held moral intuition except at its peril. But, while moral facts on the ground may enable moral practice, they do not define it.

What exactly did the "Moral Machine" study find?

Apart from, one could argue, its near philosophical irrelevance, the study reveals some very interesting things about where people in different parts of the world stand on moral issues. Three clusters representing markedly different moral priorities emerged from the data corresponding to vague regions: Western, Eastern and Southern. These groupings are not necessarily geographic (though that is a factor), nor are they circumscribed by language or historical connections (though those are factors, too).

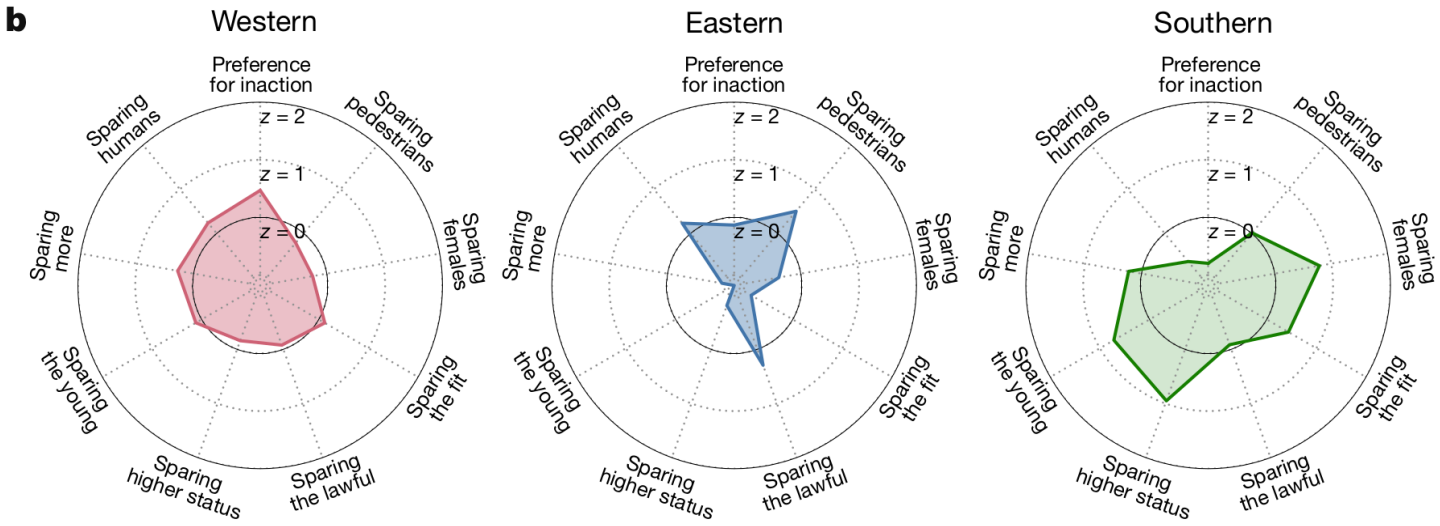
The three moral clusters sort out extensionally thus:

9. Perhaps, the mistake stems from wide recognition of the possibility of individual subjective bias. The mistake here is the inference from the unreliability of individuals to the conclusion that whole communities cannot be biased. The error might be that we falsely analogize with how scientific evidence is regarded. Where *most* of the evidence points is where we *ought* to look for "truth" in science. Evidence is what presents itself to the senses. What people say or think is not presented to the senses, rather to the "mind" for assessment. On the basis of what? On the basis of, ultimately, what is presented to the senses. *There is no special reason to take what people say as evidence of anything but what people say.* This is anthropology, background information it is good to know, but not normatively determinant.

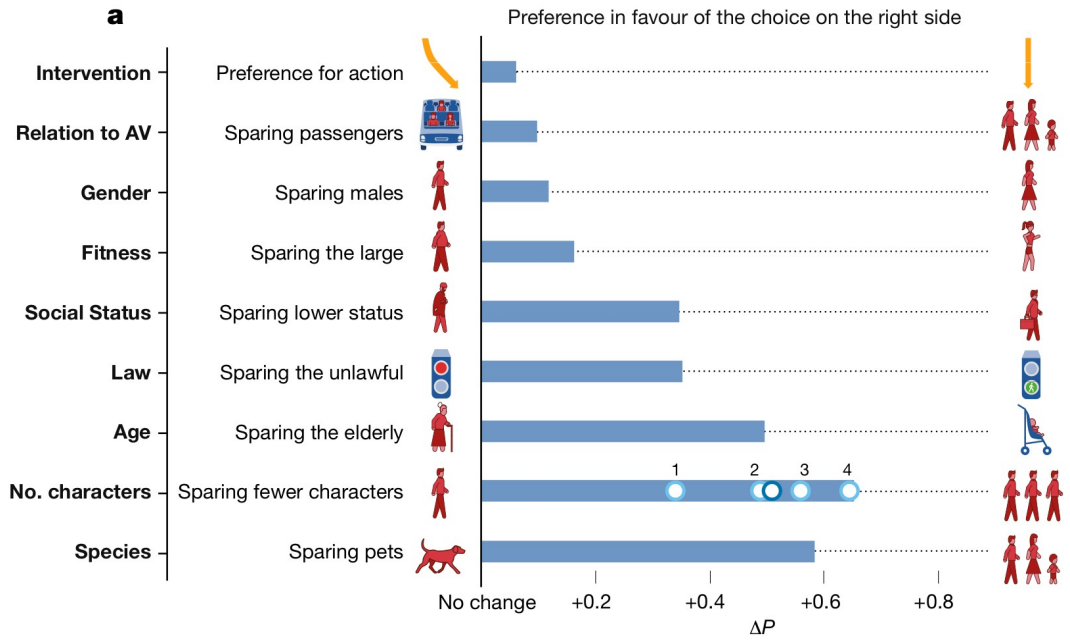


Notice that while Spain and Portugal are in the Western cluster, most of Latin America along with France and historically French-influenced cultures form the bulk of the Southern cluster. (Notice also, apparent anomalies such as Slovakia, Nicaragua, Sri Lanka, Singapore, Vietnam, and Azerbaijan/Turkey.)

Each cluster has a signature pattern of values along the nine dimensions studied:



The longer the blue bar below the more the choice on the right is favored.



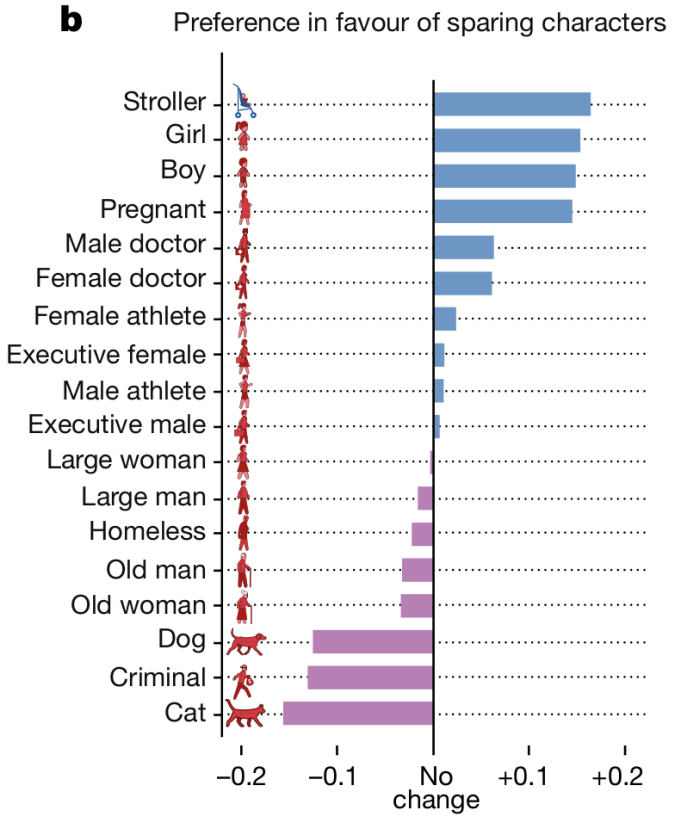
Some striking takeaways:

1. Most people are fatalists. Those in the Western cluster somewhat less so.
2. Most people are not moral egoists. At least by a little.[10] They have a *slight* preference for

10. This result is no surprise to most moral philosophers, to psychological egoists like J.S. Mill and Immanuel Kant, for example. Their respective moral theories, utilitarianism and deontology, only make sense against the background assumption that, left to their own moral "devices," those subject to these two theories, in particular, (mostly males, let's be clear) will *not* typically do the right thing. After all, Kant and Mill might argue, that's what morality is for: to compensate for this innate flaw in our nature. Ethics is a counter to natural tendencies, not an extension of them.

a car that protects *non*-passengers.

3. *Except* in the Western cluster, where we tend to expect pedestrians, all other things equal, to look out for themselves or risk death.
4. The study makes clear that females are more valuable, on the whole, than males.
5. In the Western and Eastern clusters, however, this preference for females is noticeably *less* pronounced.
6. If you are a girl, you deserve to live slightly more than a boy.
7. And, as a woman, if you are in most Southern cultures, you are *way* more valuable than a man.
8. Again, in the Southern cluster, if you are somebody's pet, you may be more valuable than your owner.



9. On the whole, it's better to be young than old if a self-driving car has you in its sights.
10. But, being young is bad news if you are in an Eastern culture where you would hardly count at all.
11. But, again, if you are in a Southern culture, being young puts you at the top.
12. Having status (socio-economic, political) will get you serious brownie points in a Southern culture. Not so much in a Western one, even less in a Eastern one.
13. "Being a homeless person is like jaywalking all the time," says Jean-Francois Bonnefon, one of the authors of the study. The value of your life is similarly discounted.

The *philosophical* takeaway from the *scientific* takeaway

So how are we supposed to program self-driving cars? The authors of the MIT study conclude that the *least* controversial values current among human societies seem to be:

1. that people count for more than animals,

Other results, further down the list, if surprising, are indications of how little we know about the varieties of normativity in human cultures.

2. that saving more lives is better than saving fewer,
3. and that the young matter more than the old.

More or less in that order. But, as noted, even about these, there is far from universal consensus. The Southern cluster places pets above people, the Eastern cluster, the elderly far above children and does not agree that more is better. And, again, these are the *least* controversial dimensions of moral ranking. The other dimensions are clearly more culturally-specific, showing less tendency to converge on anything approaching a set of universal human values.

The stated goal of the authors of the study was to inform the makers of policies that will constrain the programming of self-driving cars by offering empirical measures of what can be assumed regarding the moral values of potential buyers of such cars.

Are there universal or near-universal values that can be taken for granted in the decision-making algorithms? The answer seems to be "sort of." There are global winners (the preference for the young over the old, for example) but these global preferences have significant dissenters that are often concentrated (the Eastern cluster places noticeably *less* value on youth than the global average), making cross-community algorithmic design complex if feasible at all.

Are there major differences between moral communities? The answer is yes.

Will cars have to be designed differently for different markets? It appears so. And how many different moral communities with distinct value sets are there? The authors claim *at least* three—the Western, the Eastern, and the Southern, but there are reasons to believe that, as we make more fine-grained assessments, increasingly many and variegated subcommunities will be identified. (For example, there is reason to think that even within one country, the U.S., say, there will be regional or state differences in moral priorities, further complicating attempts at standardizing algorithms.)

Describing versus prescribing

The authors of the study (and perhaps some of their intended audience: policy makers, technologists, and manufacturers) appear to assume that ethical considerations are like other human preferences, such as those in aesthetics. One need only ask to settle questions of what to deliver. "What would you like, believe, or accept?" Although aesthetics and ethics are closely related in that they are both normative enterprises,[11] (both imply judgements and make demands),

11. Notwithstanding that some philosophers have identified ethics and aesthetics, notably Wittgenstein in the *Tractatus*, 6.421: "(Ethics and æsthetics are one.)" Wittgenstein was likely tipped off by Otto Weininger's "Logic and ethics are fundamentally the same…" in *Sex and Character*. Between the two we get: *logic = ethics = aesthetics*. They are the *only* kinds of normativity and they are fundamental to all judgment. Between them nothing escapes judgment. But Wittgenstein took himself to be remarking on something deeply subversive: namely, that one

at the level of practice, the consensus among moral philosophers is that they are radically different species of normativity. They both make pronouncements on what *should* be the case. They do not merely describe what *is* the case as science aspires to. But ethics is far more rigid and intolerant of whim and chance than aesthetics. The latter revels in *variety*. It *plays* with taste. In contrast, "ethical diversity" is a problematic concept. Ethics is prepared to righteously condemn dissent. Aesthetics may judge but "righteousness" is not on any shelf of its armamentarium. Ethics will press hard on law, for instance, to conform, whereas no legal system would justify a law because it was "beautiful" or "cool"or make something criminal literally, not metaphorically, because it was "ugly."

Consider the abortion debate. Complete indifference to, or toleration of, *both* the destruction of fetuses and the laws that seek to restrict the practice is harder to compass in the same moral framework. We want to say one side is wrong and the other is right.

If, in China, the moral consensus is that an 8 year-old's life is expendable to save an 80 year-old's, then a vehicle programmed for Latin America, where the moral trade-off is reversed, will not be acceptable.

In light of this, what are we really supposed to make of a study that purports to tell us what people *like* in different sectors of the human community? What were we expecting? That if, despite the actual results of the study, we found global unanimity of moral values, this would somehow finally answer the question what is wrong and what is right? No, it wouldn't, not in the least. It would just inform us of what humans *in fact* believe. It would tell us nothing of what they *should* believe.

If we conducted a large scale[12] study of what children would *prefer* to have for lunch, and were "shocked" to discover that they would prefer ice cream or Twinkies over fresh cut broccoli and carrots, would we take this to be an indication of what they *should* have for lunch?

But the study asked what people thought was (morally) correct in the dilemmic scenarios, not what they thought was pretty or merely desirable? So weren't moral intuitions being gauged, not aesthetic preferences?

It is not in the nature of moral feeling to see itself as needing outside justification by theory. Moral feelings are natural phenomena, which the study usefully queried. Justification for them was not part of the study. It is noticeably absent, *except* when it is suggested, as the authors of the study in fact do, that their results *should* inform public policy on regulating machine-made

could not say something about ethics that was not itself normatively neutral. Hence, all ethical theorization was tantamount to, as he put it, "gassing." Philosophers have taken pains to ignore him on this point ever since. *Wrongly*, I think, but that's another discussion...

12. For good measure.

decisions with ethical implications. The normativity in this "should" is *coming from somewhere*. It cannot come from the results of the study. These are just descriptive. The results *must* be merely descriptive or they would not be science. So then, from where? Certain assumptions made by the designers of the study? What are these? One big one is very familiar to moral philosophers: the slide from an "is" to an "ought."

Another big assumption, more specific, is that you may *derive an "ought" from an "is"* based on consequences. What consequences? Something to do with what is conducive to maximizing non-suffering states. Why those? Because, such states are conducive to survival. Why is survival held to be so fundamental? Because it is. *But why?...*

At this point, it is clear that the fundamental moral framework the authors of this study are laboring under is some form of consequentialist ethics, utilitarianism, specifically. *Why that theory?* Because—among it competitors such as deontology, virtue ethics or care ethics—it lends itself (or seems to) most easily to quantifying methods. And why would these methods be advantageous? Because they are conducive to science. And why is science so important? Because it has allowed us to progress to a level and scale of well-being far in excess of any other belief system in human experience… But we are going in circles? Haven't we already said something like this already?

The problem, in a nutshell, is that the moral theory motivating this study is front-loaded. The study does not discover anything to do with what self-describing cars *should* do. Rather it is trying predict what self-driving cars *will* do given certain facts about moral anthropology.

What's wrong with this more modest goal?

Nothing. Except that it robs some of the mystique expressed in the talk around efforts to base normativity in description. While still interesting, as any fact about humans would be, its fascination is akin to that of pornography. *So this is what it looks like to watch people making naked moral decisions...* Good. We needn't be squeamish about the facts of life. But they are just that, *facts*—which have never given us reasons to condone or condemn any course of action, human or otherwise. Those reasons, rather, come from a prior conviction that *this* or *that* is the way things *should* be. Perhaps conviction about this derives from on high (religion), from determinate or indeterminate nature (naturalist ethics), or from communal consent (contractarianism or virtue theory), or from high collective self-regard (Kantianism)… But, again, it cannot come from the data of moral anthropology.

But no such claim was made by the study. Right? If we then exclude any normative ambitions on the part of the authors of the study, what remains to be considered is more modest and less fascinating. And of limited use to self-driving car programming.

If self-driving cars ever become a reality, they are going to test our moral convictions. As long as it has been only humans with their built-in moral disabilities we have needed to consider, we might have gotten away with a lot of moral sloppiness. In a real dilemmic situation, like the one in the cartoon, we might excuse human drivers for just throwing up their hands. But AI-driven cars will not have this excuse. (They won't have hands to throw up.) AI-driven cars will not have the same excuses human drivers do because they will have the benefit of the best possible information, lightening speed reflexes, and, of course, a carefully vetted moral agenda. They will do the right thing because they will be *both* capable of it *and* programmed by the best understanding we have about what the *right* thing to do is.

Leaving aside the technical challenges, do we have an understanding of what the right thing to do is?

~

Perhaps, this is how it will go: self-driving cars are being sold to us (defended, as it were) because, when ready, they will be *safer* than human driven cars and this because they will be designed to avoid accidents better than we ever could. Will we have self-driving cars in the future? Probably. Will we have them *because* they are safer alternatives? This is not so clear. Self-driving cars are *not* especially being sold to us because they are conveniences, status symbols, or time-freeing devices, but these may well be the actual reasons why we will someday have them. Safety will *not* be our first motivation. It never has been. It will be an afterthought. No new major technological contrivance has ever been introduced and succeeded *because it was safe*. It has succeeded, if it has, because it served some *other* desire we have had. Only then, when it has more or less made itself indispensable to our ends, is there a concerted push to make it safer, to refine its negative consequences. Only then, will we be at pains to decide which moral accidents we are going tolerate.

First, many will die. *Then* we'll decide what we will do about it. This is how it has gone with commercial aviation. There is no safer way to get from one place to a distant other place than by flying. But commercial air transportation is not the paragon of safe travel it is today because it was motivated from the start by safety. It was motivated by things like efficiency, status symbolism, or by freeing travelers from the tedium of traversing long distances. These then became *de rigueur*. Once so, great effort was made to cut down on the carnage... Since we are not moral or rational, despite our talk—since we are not really motivated by what we say we are, why expect our machines to be?

There will be self-driving cars someday, but we won't be bothered to buy them. Instead, we will allow government to force them on us if we know what is good for us. For reasons similar to why the IRS does not use the honor system.

We want incompatible things and we want them together

What do we want from a mode of transport?

1. *Utility:* it gets us where we want to go
2. *Convenience:* it's available without unwanted expenditures of effort.
3. *Safety:* it doesn't harm us or anyone or anything we care about.
4. *Enjoyment:* (perhaps) self-explanatory.

The first three are services best supplied as a public utility. (Not necessarily a utility we have, but one it is conceivable we might have.) We don't, as individuals, need to *own* such transport. In fact, it's not clear what point there would be in our *individual* ownership.

However, the last of these desires may require ownership because *enjoyment in moving* derives essentially from autonomy, not the conveyance's autonomy, but *our* autonomy if we conceive of autonomy as personal and physical: our legs, a bicycle, or a motorized conveyance whose direction and speed we control with our limbs. *This is visceral autonomy and it is what is enjoyable about moving.* An essential element of this enjoyment is risk. However slight, the risk must be real. We want to know that if we don't watch where we are going we may hurt ourselves, others, or something. When we are in the mood to do it, it's *fun* to watch where we are going.

But, if all we truly want is to get from A to B, without regard to any particular experience involved in getting there, whether because we don't want to be bothered with being careful or we have better things to occupy our concentration, ownership of the conveyance, either literal ownership or ownership in the form of autonomy (because autonomy is intimately connected with ownership)—*if that's all we want, there is nothing very interesting or useful about a self-driving car.* Why should we care whether the conveyance drives itself or has a remote pilot? Its automation is as interesting as the automation of a toaster. So long as it fulfills those first three requirements: it gets us where we want to go conveniently and safely.

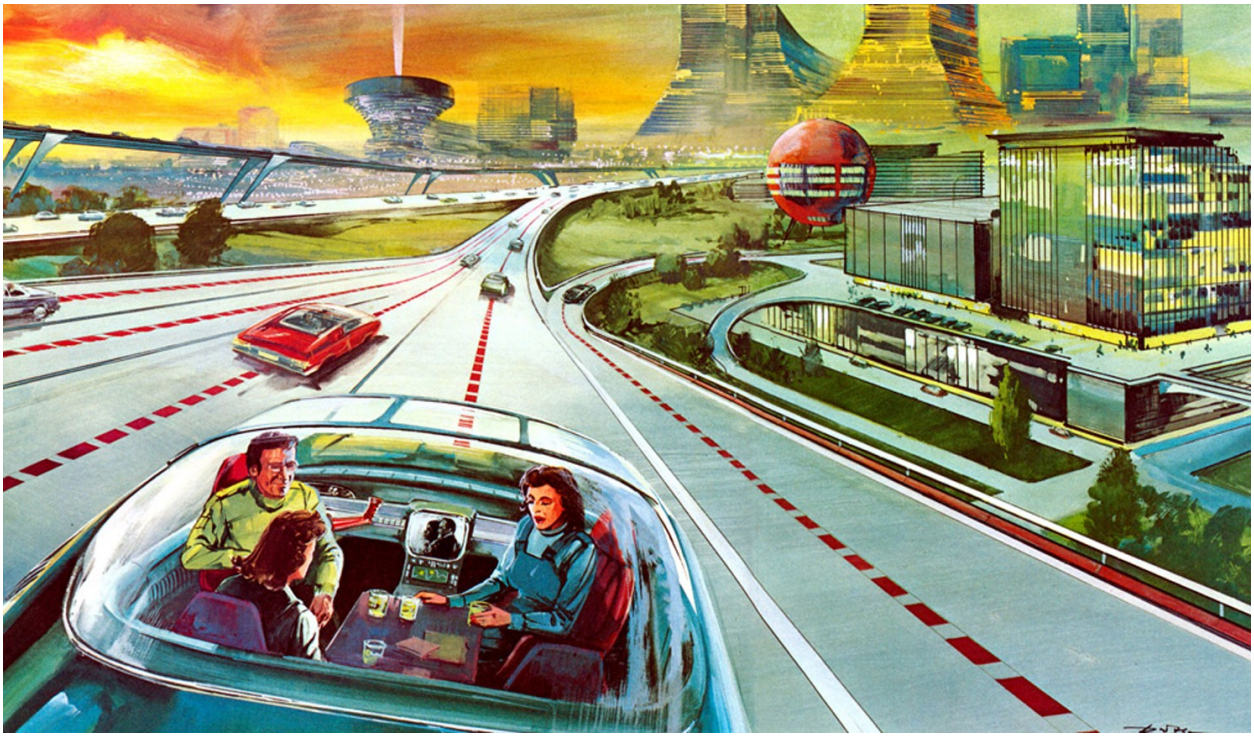
The concepts of *autonomy* and *safety* do not cohere. Why would I want to *own* a vehicle that was optimized for safety? It would be an absolute bore. If I want to own it, it's because I want to *own* what it does. And that necessarily entails risk.

~~

Part 4 of this series on AI and philosophy will explore AI governance: Since we have no track

record of governing ourselves either effectively or humanely, why not turn this work over to intelligences less freighted with our nasty evolutionary legacies?

How about it? *AI-governed nations?* (Yes, there is some high-powered thinking going into this.)



Resources

Literature

"Is the Trolley Problem Derailing the Ethics of Self-Driving Cars? Pacific Standard spoke with a philosopher who's trying to code ethical algorithms into autonomous vehicles." Jack Denton *Pacific Standard*, Nov 29, 2018.

"How should autonomous vehicles be programmed? Massive global survey reveals ethics preferences and regional differences." Peter Dizikes, *MIT News Office*, October 24, 2018.

"The Moral Machine experiment," (the aforementioned "massive survey") Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon & Iyad Rahwan. *Nature*, volume 563, pages 59–64 (2018).

"Ethics, algorithms and self-driving cars—a CSI of the 'trolley problem'," Andrea Renda, *Policy Insights*, January 2018. Renda offers a "crime scene investigation." What kind of

questions are we going to ask of the car?

"The ethics of crashes with self-driving cars: A roadmap, I," Sven Nyholm, *Philosophy Compass*, Volume 13, Issue 7, July 2018.

"Why Trolley Problems Matter for the Ethics of Automated Vehicles," Geoff Keeling, *Science and Engineering Ethics*, March 2019.

"If Buddhist Monks Trained AI: Monks take on the trolley problem, a classic moral dilemma that has big implications for driverless cars," Alexis C. Madrigal *The Atlantic*, Jun 29, 2017. Are Buddhist monks psychopaths or what? Or just more sensible than the rest of us?

"'The Trolley Problem' in Judaism: Sacrifice One Life to Save Many?" Yehuda Shurpin, Chabad.org. Jews have had the trolley problem licked since ancient times. Everybody's blood is red. The blood of five is no redder than the blood of one. Let God decide. But there are exceptions! Bad hombres are not equal to regular hombres. (Thanks to Mary for this link.)

Podcasts

"Episode 30: Derek Leben on Ethics for Robots and Artificial Intelligences," (Podcast) Philosopher Derek Leben is interviewed at length by Sean Carroll. This highly recommended discussion pinpoints what is at stake for moral theory.

"Episode #40 – Nyholm on Accident Algorithms and the Ethics of Self-Driving Cars," (Podcast) Philosopher Sven Nyholm is interviewed by John Danaher June 29, 2018.

Video

"Moral Machines: How culture changes values," (Short video describing the Moral Machine experiment.) How do different cultures value human life? To find out, researchers created a viral online experiment to gather data from millions of participants across the world. Some values generalised across cultures, but others came as a surprise.

Jean-Francois Bonnefon, Toulouse School of Economics, held a keynote, "The Moral Machine Experiment", (longer video describing experiment) at IJCAI-ECAI 2018, the 27th International Joint Conference on Artificial Intelligence and the 23rd European Conference on Artificial Intelligence, the premier international gathering of researchers in AI.

"The Social Dilemma Of Driverless Cars | Iyad Rahwan | TEDxCambridge," Iyad Rahwan, associate professor at the MIT Media Lab, explores how Artificial Intelligence challenges our morality. If a driverless car must choose between killing five pedestrians crossing the street, or swerving to a wall and killing its own passenger, what should it do? Decide for yourself at http://moralmachine.mit.edu.

"MIT Self-Driving Cars: State of the Art (2019)," Lex Fridman. Introductory lecture of the MIT Self-Driving Cars series with an overview of the autonomous vehicle industry in 2018 and looking forward to 2019, including Waymo, Tesla, Cruise, Ford, GM, and out-of-the-box ideas of boring tunnels, flying cars, connected vehicles, and more. The talk covers the state of the art in terms of industry developments.

Notes

From Renda's paper:

> Other academics use game theory to conclude that pedestrians would impose their presence on risk-averse self-driving cars, especially if they behave with impunity. Scholars imagine cars equipped with an "ethical knob", which could set key patterns of behaviour such as "full altruist", "fully egoist", or "impartial". Others opt for simple "value of life" models as opposed to neural networks, since the latter are "black boxes", and as such would be largely unacceptable to end users….
>
> Market players and regulators have not been silent either. Last year the German car manufacturer Mercedes took an explicit stance on this ongoing *querelle* by announcing that indeed, their cars were being designed to prioritise occupants over pedestrians. [p. 1]

...

In the film *Sophie's Choice*, why did she give up the girl and not the boy? There's a reason. It wasn't random. (Hint: if she had been the father instead of the mother, the *right* choice would have been different.)

…

Rogue Boeing 737 Max planes 'with minds of their own' | 60 Minutes Australia

https://youtu.be/QytfYyHmxtc

[Part 3 of robot ethics: parts 1 (consciousness and suffering) and 2 (robot love).]